Scalable Sequence Analysis with FastHMM, FastBLAST, and FastTree

Morgan N. Price[1,2], Paramvir S. Dehal[1,2], Adam P. Arkin[1,2,3]

[1]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov/; [2]Lawrence Berkeley National Laboratory, Berkeley, CA, 94720; and [3]Department of Bioengineering, University of California, Berkeley, CA, 94720

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy

As DNA sequencing accelerates and the sequence databases grow, many problems in sequence analysis are becoming computationally challenging. We have developed fast heuristics for placing sequences into known protein families, for identifying new protein families, for identifying all homologs of a protein, for inferring a phylogeny for a gene family, and for identifying orthologs. FastHMM and FastBLAST identified domain assignments and homology relationships for 6.5 million proteins from the non-redundant Genbank database ("NR") in just 14,000 CPU-hours instead of an estimated 380,000 CPU-hours for HMMer 2 and all-versus-all protein BLAST. FastTree inferred a phylogeny for the largest known sequence family, the 16S ribosomal RNA family, with 212,000 distinct sequences, in just 48 hours (with 1 CPU) and 5.5 GB of RAM. We are not aware of any other practical method that can construct a tree for such large families. Finally, given the homology relationships and the phylogenetic trees for families, orthologs can be identified quickly. These methods are integrated into the MicrobesOnline web site, and together, these ensure that MicrobesOnline will scale to handle thousands of genomes.